Layer 2 Leaf & Spine Design and Deployment Guide

The intention of this guide is to provide a systematic and well thought out series of steps to assist the reader with the design and deployment of a Layer 2 Leaf and Spine (L2LS) topology. The example deployment is based on a design which meets a set of predefined requirements as listed in the System Requirements section of this guide. This guide was written for network engineers, administrators and cloud architects but can be informative to any role dealing with layer 2 networking. A good working knowledge of networking principles is assumed.

The guide is broken down into four main sections as noted below, each section begins with a high level overview and gets progressively more detailed.

- System Requirements
- Design Overview
- Detailed Design
- Configuration Examples

Table of contents

CONTENTS

The Drivers for Layer 2 Leaf + Spine Topologies

Legacy datacenters are frequently challenged by increasing amounts of East to West traffic. The main architectural reasoning for this is legacy datacenters were designed with a 3-tier model consisting of core switches, distribution/aggregation switches and access switches. Increasing port density, which merchant silicon has greatly improved, allows for more modern designs that allow architects to eliminate the distribution/aggregation layer and enable what is now called a Leaf and Spine architecture.

One of the primary decision points when building a Leaf and Spine topology is whether the links between the Leaf and Spine switches should be Layer 3 or Layer 2 links. This is commonly driven by the application or workload but in either case Layer 2 connectivity on at least some segments is typically required. Although Layer 2 connectivity can be accomplished on a Layer 3 topology by utilizing an overlay technology such as VXLAN, some organizations do not require the scale offered by building a Layer 3 underlay and a Layer 2 topology utilizing technologies like Multi-Chassis Link Aggregation (MLAG) is sufficient.

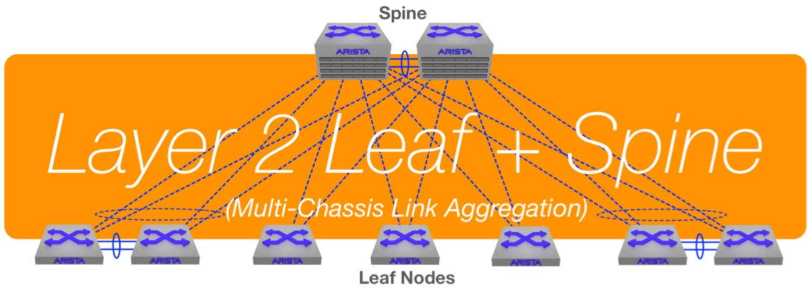

Figure 1: Layer 2 Leaf and Spine Topology

Another constraint legacy datacenters have suffered from is providing active/active connectivity to all switches. To overcome this constraint Multi-Chassis Link Aggregation is utilized (MLAG). MLAG allows architects to design resilient and high performing networks by ensuring that every link is active and spanning tree is not necessary on the links between the Leaf and Spine. This design simulates a Clos topology. The Clos network was originally envisioned by Charles Clos in 1952 and was based on circuit switching technologies. Arista has pioneered the modern day Clos network by leveraging Ethernet, IP protocols and high density switching silicon.

In summary, leveraging Ethernet technology and standard protocols to build redundant and resilient L2LS networks provides the best of both worlds and is the foundation for Arista's Universal Cloud Network (UCN) Architecture.

System Requirements

The table below details a list of typical requirements seen in real datacenter specifications. The goal of this guide is to ensure all of these requirements are met as well as demonstrate the necessary system configurations to deploy them. The requirements include aspects of network and server requirements.

System Requirements	
Spine Redundancy	There is requirement to have two spine/core switches to share the load.
Spine Resiliency	The requirement is to have the ability to remove a spine switch from service or suffer a spine failure and have it return to service with little or no impact on application traffic flow.
Scalability	The network must be able to seamlessly scale to support future bandwidth requirements.
Non-Blocking	Design must support the ability to implement leaf-spine subscription ratios based on specific application requirements.
Congestion Avoidance	Design must have the capability to absorb peak traffic loads without losing packets. The network must have mechanisms to ensure traffic can be prioritized and queued if necessary to eliminate the potential of packet loss.
Active/Active Server Connectivity	Each server must be able to have active/active connections, one to a primary leaf switch and one to secondary leaf switch.
Open Standards	The network must support open standards based protocols, no vendor proprietary protocols or features will be used.
Edge Connectivity	Network design to include connectivity into the LAN / WAN environment.
Network Address Translation	Native server IP's must be hidden from the outside network i.e.) Network Address Translation (NAT) must be supported at the network edge.
Traffic Engineering	Mechanisms to ensure traffic can be prioritized and or dropped based on policies.

Arista Universal Cloud Network (UCN) Architecture

The system requirements outlined in the previous section can be met with Arista's Universal Cloud Network (UCN) Architecture. The diagram in Figure 1 depicts the components that make up this architecture. For a comprehensive overview of Arista's UCN architecture you can download the Arista Universal Cloud Network Design Guide at http://arsta.co/2mGK3C1

The focus of this guide is the Layer 2 Leaf and Spine (L2LS) and specifically the design, components and configuration details required to build, test and deploy it.

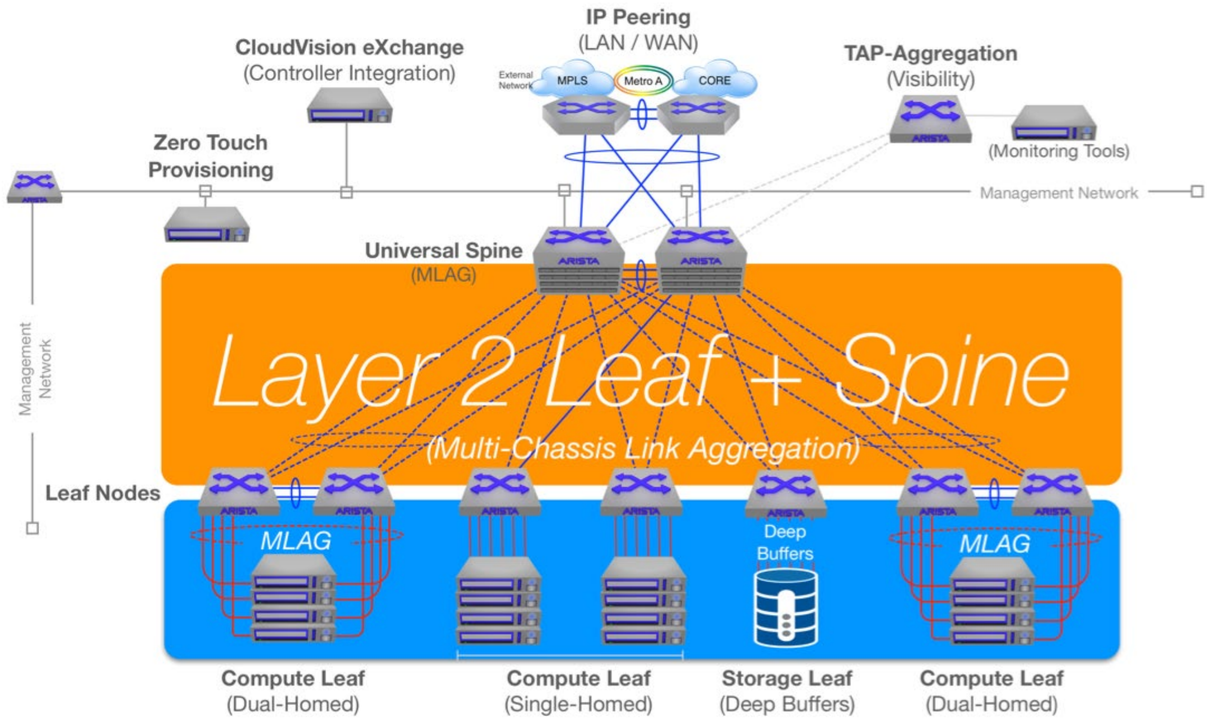

Figure 2: Arista's Universal Cloud Network (UCN) Architecture

Design Overview

The Layer 2 Leaf and Spine (L2LS) topology is the foundation of Arista's Universal Cloud Network Architecture. At a high level, the Layer 2 Leaf Spine (L2LS) is a 2-Tier topology comprised of spine and leaf switches. This simple design, when coupled with the advancements in chip technology, a modern operating system and an Open standards approach, provides significant performance and operational improvements.

One of the main advantages of the L2LS design is that the pair of spine switches are presented to the leaf-layer switches as a single switch through the use of MLAG (Multi-chassis Link Aggregation Group), which inherently allows LAYER 2flexibility throughout the environment. This also eliminates the dependence on spanning-tree for loop prevention and allows for full utilization of all links between the leaf and spine. It is worth noting that the scalability of the spine is limited to a total of 2 switches.

By adopting a merchant silicon approach to switch design, architects are now able to design networks that have predictable traffic patterns, low latency, and minimal oversubscription. Legacy designs often incorporated more than two tiers to overcome density and oversubscription limitations.

Leaf and spine switches are interconnected with LAG (802.3ad) links and each leaf has at least one connection to each spine.

A Universal Spine

Arista believes the data center network spine should be universal in nature. What this means is that by using standard protocols and design methods coupled with robust hardware components the data center spine can be leveraged throughout the campus.

The concept behind Arista's Universal Spine is:

- Build a spine that meets the needs of all Data Centers
- Reliable and simple to operate
- Interoperate with any and all vendors' equipment, leaf switches, firewalls, Application Delivery Controllers etc.

A two-node spine is depicted in Figure 3 and will be used as the working example for this guide.

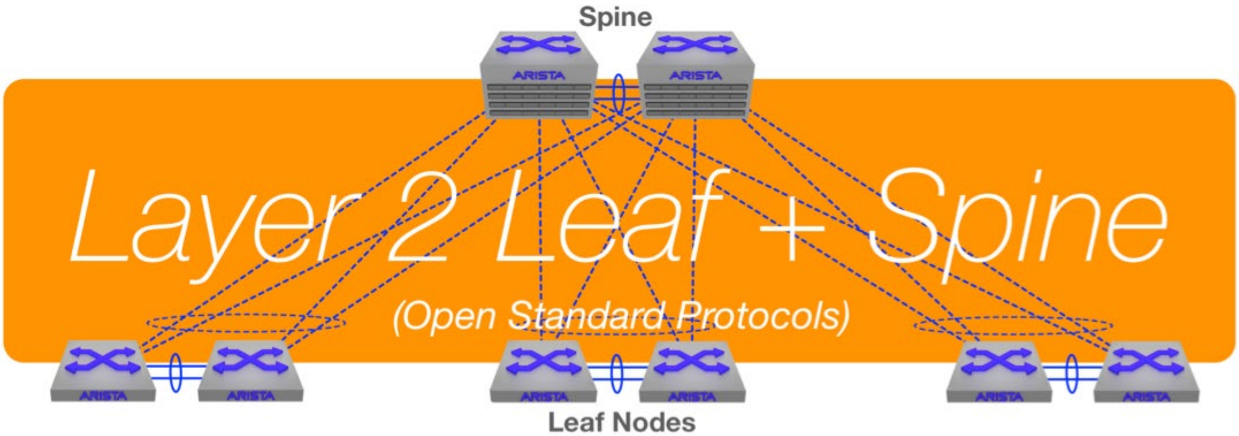

Figure 3: Arista Layer 2 Leaf & Spine (L2LS)

Multi-Chassis Link Aggregation

Multi-Chassis Link Aggregation Group (MLAG) is a standards based approach to providing active/active link aggregation across two switches. Traditional redundant Layer 2 path designs leveraged spanning-tree protocol (STP) to provide a loop-free topology through the blocking of links. This approach while providing the redundancy reduced the value of investing in high-speed links and was often fraught with slow convergence. MLAG changes this by providing full value for the investment in redundant paths, reduces complexity and reduces convergence speeds associated with failures in layer 2 domains.

A MLAG is a bundle of links that terminate across two physical switches and appear to the downstream device as a single link aggregation group (LAG). The switches terminating the MLAG form a peer adjacency across a peer link over which state is shared and a single MLAG System ID identifies the MLAG pair. The connected layer 2 device sees a single logical STP bridge or LACP node. As the MLAG pair is seen as a single STP bridge, all links are forwarding and there is no requirement for blocked links.

In the L2LS design, MLAG is used in the spine between two switches. It is also an option for the leaf nodes to provide redundant connectivity to hosts.

Leaf Options

Leaf switches ensure compute, storage and other workloads get the necessary connectivity and bandwidth dictated by the applications they serve. In the past the inter-connections between the leaf and spine switches have been heavily oversubscribed, careful consideration needs to be taken to ensure the appropriate amount of bandwidth is provisioned. With greater server densities, both virtual and physical, port densities and table sizes are another consideration that needs to be taken into account when selecting platforms.

From a leaf design perspective there are two main configurations that need to be considered, designs that support Single-Homed workloads and Dual-Homed workloads as well as a workload such as IP Storage that may benefit from a deep buffer solution such as a Storage Leaf.

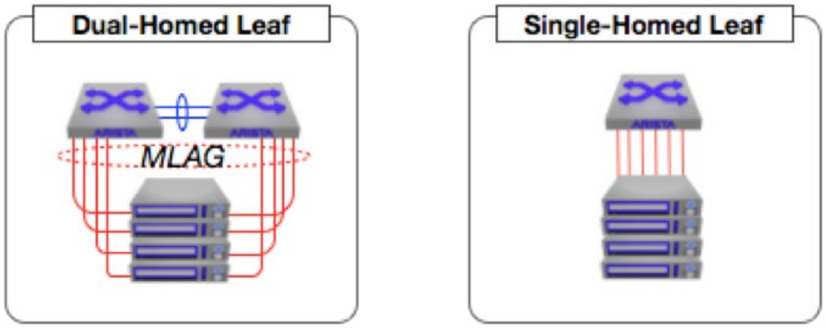

Figure 4: Single-Homed and Dual-Homed Leaf Configurations

Dual-homed systems will leverage a MLAG (Multi-Chassis Link Aggregation) configuration. The MLAG configuration supports the requirement for active/active server connectivity with switch level redundancy at the Top of Rack (ToR).

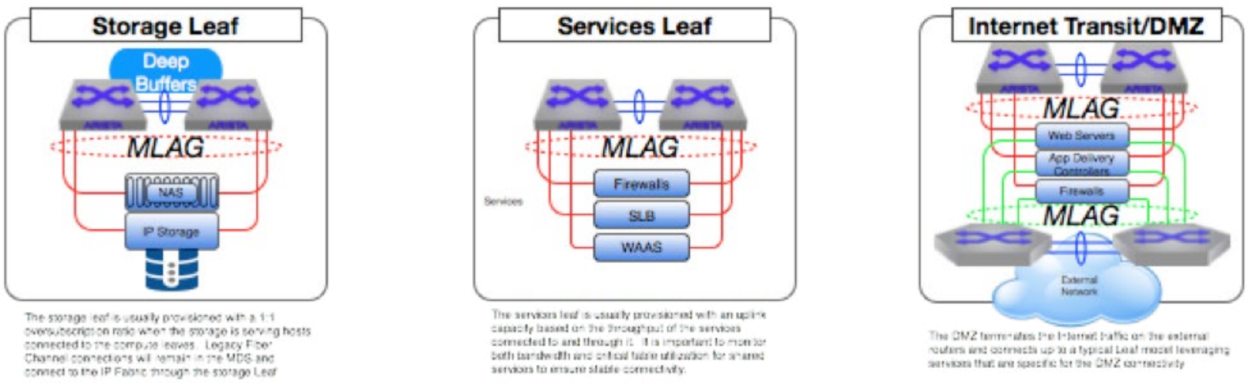

Figure 5: Storage, Services and Internet DMZ Leaf Configurations

Dedicated Storage Leafs can also be provisioned to ensure IP based storage systems can endure the sustained traffic bursts and heavy incast events. In this design a deep buffer switch should be utilized near the storage system to support the requirements. Deep buffers ensure fairness to all flows during periods of moderate and heavy congestion.

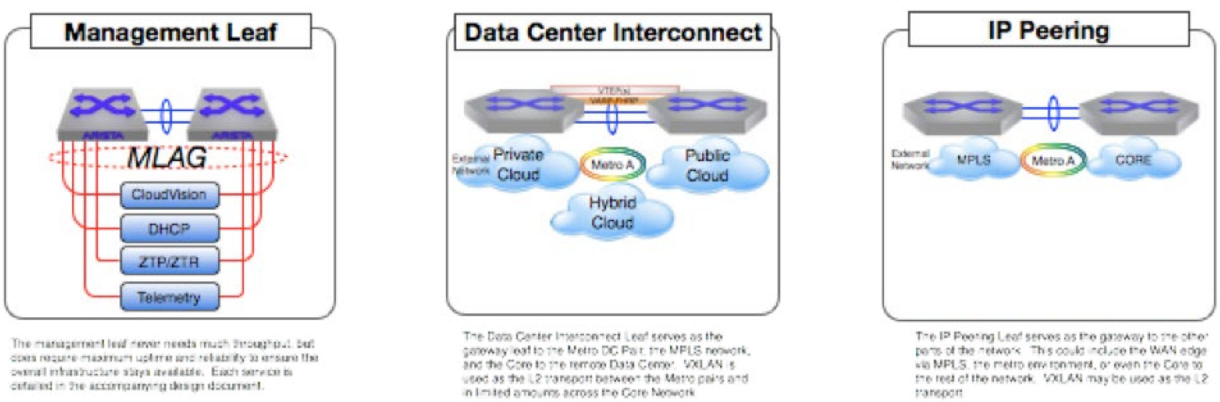

Figure 6: Management, Data Center Interconnect and Edge/Border Leaf Configurations

Regardless of the leaf configuration, single-homed, dual-home or otherwise, the leaf to spine connectivity is consistent ensuring a common configuration throughout the fabric. Arista has a number of spine and leaf platforms to choose from that meet a broad range of requirements. A more detailed look at the Arista Cloud Network Portfolio can be found at http://www.arista.com/en/products/switches.

LACP

A MLAG pair at the leaf provides an option for dual-homing server connections. The server does not require knowledge of MLAG and can simply be configured with dynamic or static LACP. This provides for an easy to deploy and standards based approach to providing redundant server connections.

Spanning-Tree

An Ethernet network functions properly when only one active path exists between any two stations. A spanning tree is a loop-free subset of a network topology. STP is a LAYER 2network protocol that ensures a loop-free topology for any bridged Ethernet LAN. Spanning-Tree Protocol allows a network to include spare links as automatic backup paths that are available when an active link fails without creating loops or requiring manual intervention. The original STP is standardized as IEEE 802.1D.

Several variations to the original STP improve performance and add capacity. Arista switches support these STP versions:

- Rapid Spanning Tree (RSTP)
- Multiple Spanning Tree (MSTP)
- Rapid Per-VLAN Spanning Tree (Rapid-PVST)

RSTP is specified in 802.1w and supersedes STP. RSTP provides rapid convergence after network topology changes. RSTP provides a single spanning tree instance for the entire network, similar to STP. Standard 802.1D-2004 incorporates RSTP and obsoletes STP.

The RSTP instance is the base unit of MST and Rapid-PVST spanning trees.

Rapid Per-VLAN Spanning Tree (PVST) extends the original STP to support a spanning tree instance on each VLAN in the network. The quantity of PVST instances in a network equals the number of configured VLANs, up to a maximum of 4094 instances.

Multiple Spanning Tree Protocol (MST) extends rapid spanning tree protocol (RSTP) to support multiple spanning tree instances on a network, but is still compatible with RSTP. By default, Arista switches use MSTP, due to it's increased convergence and scale out capability as compared to PVST and RSTP.

MST supports multiple spanning tree instances, similar to Rapid PVST. However, MST associates an instance with multiple VLANs. This architecture supports load balancing by providing multiple forwarding paths for data traffic. Network fault tolerance is improved because failures in one instance do not affect other instances.

First Hop Redundancy Protocols

In a L2LS design, the spine MLAG pair is the logical place for the layer 3 gateways. MLAG works well with both Virtual Redundant Router Protocol (VRRP) and Virtual ARP (VARP). VARP is the preferred First Hop Redundancy Protocol (FHRP) in the L2LS design and is detailed further in this section.

VARP

Virtual ARP (VARP) allows multiple switches to simultaneously route packets from a common IP address in an active/active router configuration. Each switch in an MLAG pair is configured with the same set of virtual IP addresses on corresponding VLAN interfaces and a common virtual MAC address. In MLAG configurations, VARP is preferred over VRRP because VARP does not require traffic to traverse the peer-link to the master router as VRRP would.

VARP functions by having each switch respond to ARP and GARP requests for the configured router IP address with the virtual MAC address. The virtual MAC address is only for inbound packets and never used in the source field of outbound packets.

When IP routing is enabled on the switch pair, packets to the virtual MAC address are routed to the next hop destination.

Multicast

Layer 2 Multicast

Leaf switches ensure compute, storage and other workloads get the necessary connectivity and bandwidth dictated by the applications they serve. In the past the inter-connections between the leaf and spine switches have been heavily oversubscribed, careful consideration needs to be taken to ensure the appropriate amount of bandwidth is provisioned. With greater server densities, both virtual and physical, port densities and table sizes are another consideration that need to be taken into account when selecting platforms.

Layer 3 Multicast

In the L2LS design, the MLAG pair of spine switches running VARP for redundant layer 3 gateways may also be configured as the layer 3 multicast routers. Protocol Independent Multicast (PIM) interoperates with VARP to provide an active/active multicast gateway.

The PIM design in an MLAG pair works as follows:

- Both MLAG peers are configured and active PIM routers
- Both peers are capable of generating PIM frames and processing IGMP joins
- No synchronization of state or configuration is necessary between the peers
- Peers operate independently, building their own MFIB

IGMP SNOOPING

IGMP Snooping "snoops" (or listens) to the IGMP reports being sent from a host to a multicast router. The switch listens to these reports and records the multicast group's MAC address and the switch port upon which the IGMP report was received. This allows the switch to learn which ports actually need the multicast traffic, and will send it only to those particular ports instead of flooding the traffic.

IGMP SNOOPING QUERIER

Layer 2 only VLANs, or those without locally configured PIM, will rely on the existence of an IGMP snooping querier to prevent flooding of multicast traffic, and for the IGMP snooping feature to properly learn and program host ports which should be receiving multicast traffic.

The IGMP querier feature runs automatically when PIM is configured on the local SVI (layer 3 VLAN interface). If for some reason, the network design does not allow for PIM to be configured anywhere (which often is the case in networks where administrators do not wish to allow multicast to cross between VLANs), Arista switches allow for a global, or per vlan, option to enable the IGMP querier feature.

IGMP SNOOPING MROUTER

Layer 2 switches that receive IGMP membership reports from connected hosts need to know where to forward these control packets. An mrouter port is the upstream interface on a switch, which points towards the Layer 3 multicast PIM router. Mrouter ports are dynamically learned when PIM hello messages are received by the switch.

When IGMP membership reports are received by a layer 2 switch, the switch checks to see if there are any dynamically learned mrouter interfaces, and forwards a copy of these reports out of the mrouter port(s).

The purpose of this is to inform the local PIM router(s) that there are interested hosts on the various LAYER 2segments that want to receive multicast traffic. PIM can then update the associated state accordingly for any (S,G) entries to forward or prune.

Link Layer Discovery Protocol

The Link Layer Discovery Protocol (LLDP) is a vendor-neutral link layer protocol used by devices for advertising their identity, capabilities, and neighbors on a local area network.

As data centers become more automated LLDP is being leveraged by many applications as a way to automatically learn about adjacent devices. Adjacent is a key word as LLDP only works between devices that are connected at layer two i.e.) in a common VLAN.

At the network level LLDP becomes important when integrating with provisioning systems such as OpenStack and VMware. Through LLDP, switches are able to learn details about connected devices such physical / virtual machines, hypervisors as well as neighboring switches. As an example of this type of integration, when VM instances are created on compute nodes the Ethernet trunk port between the leaf switch and compute node can be automatically configured to allow the required VLANs, this is enabled by using information learned by LLDP.

The use of LLDP should be considered and reviewed with virtualization and cloud architects during the design activities.

Detailed Design

The L2LS has a number of elements that need to be considered during the detailed design. At a high level this work can be broken down into Leaf and Spine.

The diagram below begins to reveal some of the finer points of the layer 2 Leaf and Spine design. For the purpose of this exercise some assumptions will be made about many of these details however it will be noted and explained as to how and why these design choices could apply to network planning. The stated system requirements will also guide our decision-making. The design includes 100G between leaf and spine, though 40G is also an option depending on requirements.

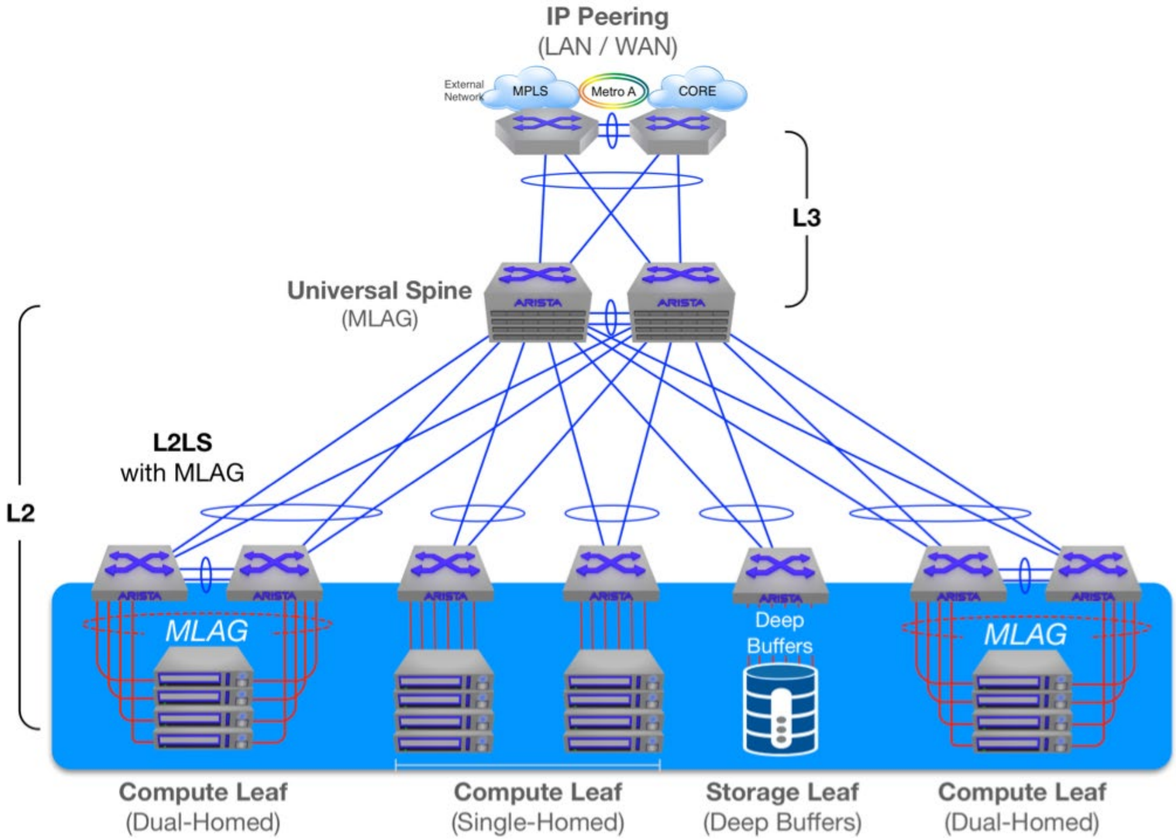

Figure 7: Layer 2 Leaf/Spine with MLAG (2-Way Spine)

Leaf Design Considerations

Application and server requirements dictate leaf design choices and platform selection. The Leaf Design section is broken as follows: Interfaces and Port Densities, Subscription Ratios, Table Sizes, Single and Dual-Home Workloads, Layer 2 Requirements and Transceivers and Cabling.

Interfaces and Port Densities

One of the first things that need to be considered when beginning design work is what type of workload will be supported and how many workloads will need to be scaled to. Getting detailed answers to these questions upfront is paramount to making good design decisions.

There are a number of interface choices for physical server connectivity today and the list is growing. Below is a list of interfaces that require consideration. The list is not a comprehensive of all interface types but focuses on short-range optics commonly seen with the data center.

- Small Form-factor Pluggable (SFP): 1GBASE-T, 1GBASE-SX, 10GBASE-T, 10GBASE-SR
- Quad Small Form-factor Pluggable (QSFP): 40GBASE-SR
- Multi Speed Ports (MXP): 10, 40 & 100G
- QSFP100: 25GBASE-CR, 50GBASE-CR, 100GBASE-SR4

Teams outside of the network-engineering group may drive interface requirements at the end of the day. With traditional DC networks this was less of a concern however new interface types and speeds have changed several things. The first one being parallel lanes and the second being cabling types such as MTP, in both cases it requires a good understanding of optical requirements, specifically 40G but also 25G, 50G and 100G.

There are also situations where retrofitting or upgrading an existing data center network is necessary, which leaves engineers in a situation where they are forced to adopt existing cabling plants.

The quantity of servers within each rack and the anticipated growth rate will need to be documented. This will dictate the switch port density required, Arista has a broad range of platforms in densities from 32x10GBaseT ports (with 4x40G uplinks) in a 1RU design all the way to 64x100G ports in a 2RU leaf platform with many variants in between.

Arista’s Portfolio of Cloud Networking switches can be found https://www.arista.com/en/products/switches.

Transceivers and Cables

As with any network design transceiver and cabling types need to be determined up front. A deep dive on cabling and transceiver selection is beyond the scope of this guide however more information can found at https://www.arista.com/en/products/transceivers-cables.

For this design exercise 40G uplinks have already been determined however the cabling/media has not. If runs are short enough Twinax or Direct Attached Cables can be used and are very cost effective. Active Optical Cables (AOC) are another good economical choice, both of these cable types have integrated QSFP transceivers. For an existing Multi-Mode (MM) or Single-Mode (SM) fiber plant there are a number of choices. For this design guide a MM plant will be used.

Leaf Uplinks

The type and number of uplinks required at the leaf switch is dictated by the bandwidth and spine redundancy/resiliency needed. With a four-node spine for example a minimum of four 10G or 40G uplinks would be utilized.

Traffic Load Balancing

Switches balance packet load across multiple links in a port channel by calculating a hash value based on packet header fields. The hash value determines the active member link through which the packet is transmitted. This method, in addition to balancing the load in the LAG, ensures that all packets in a data stream follow the same network path.

In network topologies that include MLAGs or multiple paths with equal cost (ECMP), programming all switches to perform the same hash calculation increases the risk of hash polarization, which leads to uneven load distribution among LAG and MLAG member links. This uneven distribution is avoided by performing different hash calculations on each switch in the path.

Available hashing methods on Arista switches are:

- MAC header (dst-mac, src-mac, eth-type, vlan-id, vlan-priority)
- IP fields (dscp, dst-ip, src-ip, protocol, src-port, dst-port)

Table Sizes

With the adoption of virtualization technologies there has been an explosion of virtual machines on the network. In a somewhat silent manner these virtual machines have increased the MAC addresses count on the network significantly. Packet processors have finite resources that need to be considered during the design. It is important to understand the scaling requirements for MAC addresses, ARP entries and route tables in order to make the appropriate platform selections and place layer 3 boundaries accordingly.

Single-Homed Workloads

Certain applications are designed in a fault tolerant manner to support hosts joining and leaving the workload dynamically. Such workloads can be attached to a single leaf and rely on the underlying application for redundancy rather than the network. Single-homed hosts can be connected to the same leaf switches as dual-homed workloads as long as sufficient bandwidth is provisioned on the MLAG peer link.

Dual-Homed Workloads

Another requirement for this design includes network level fault tolerance for server connectivity. Network level fault tolerance is the ability for a workload to survive a single switch failure (at the rack level) without impacting host connectivity and ultimately application availability. Generally speaking network level fault tolerance assumes active/active server connectivity.

To support these requirements a Dual-Home Leaf Configuration utilizing MLAG will be used. MLAG is standards based and is interoperable with any device that supports the Link Aggregation Control Protocol (LACP) / 802.3ad specification. This configuration supports fault tolerance and active/active load sharing for physical and virtual servers.

Border Leaf

The Border Leaf provides connectivity to resources outside of the Leaf and Spine topology. This may include services such as routers, firewalls, load balancers and other resources. All though the Border Leaf is deployed in a similar manner as other leafs switches, traffic traversing the Border Leaf is typically considered to be North-South traffic rather than East-West. The Border Leaf requires specific consideration as it is often connected to upstream routers at a fraction of speeds of the network spine. For example, a typical leaf-spine interconnect would be running at 40G and a Border Leaf could be connected to the outside world at 1 or 10G. This speed change needs to be understood as higher speed links can easily overwhelm lower speed links, especially during burst events. Routing scale and features also need to be taken into account.

Edge Connectivity

To connect the Border Leaf to the Edge Routers, a peering relationship must be established. From a protocol perspective, BGP provides the best functionality and convergence for this application. Whether to use iBGP or eBGP truly depends on the specific requirements and capabilities of the edge devices. External BGP (eBGP) is primarily used to connect different autonomous systems and as such lends itself well to this design scenario whereby a Border Leaf, representing the data centers autonomous needs to connect to the Edge Router, which may represent any number of different autonomous systems.

Depending on the use case and specific design requirements, attention must also be given to how public prefixes will be shared with the Edge Routers and how access to the Internet will be routed. If this is a requirement, the Border Leaf must be prepared to accept and redistribute a default route to the spine. Conversely, any private autonomous system numbers (ASN) must be removed before advertising to upstream providers and transit peers, this can be done in a number of ways and is beyond the scope of this document.

Spine Design Considerations

Characteristics of a Network Spine

As a subset of the initial requirements presented in this guide the network spine requires careful consideration. The network spine is expected to be robust and support all types of workloads at both low and peak loads. A proper spine design should be able to scale as the business grows without the need for forklift upgrades and have a 5+ year lifespan. Last but not least the spine should provide deep visibility into switch performance data while at the same time be easy to update and automate.

Key Spine Attributes

There are also several more specific attributes that require consideration. In larger networks or networks built with chassis based systems the design needs to take into consideration the Internal Switching Fabric itself. Switching fabrics can be broken down into two main categories, Ethernet/Flow-Based and Cell-Based. Much the same as leaf design, queuing and buffering are also considerations as is the tolerance level to accept packet loss in the network. Table sizes, power and density as well as cost are always considerations as well. Using open standards based protocols are also key attributes of any good design.

Internal Switching Fabrics

In chassis based systems (as well as multi-chip systems) their needs to be a way to connect front panel ports from one line card to ports on other linecards. These connections are made behind the scenes via specific fabric packet processors or other types of internal switching chips. The key take away here is that there is more than one type of "internal switching fabric" and it is important to understand the differences when making spine design decisions.

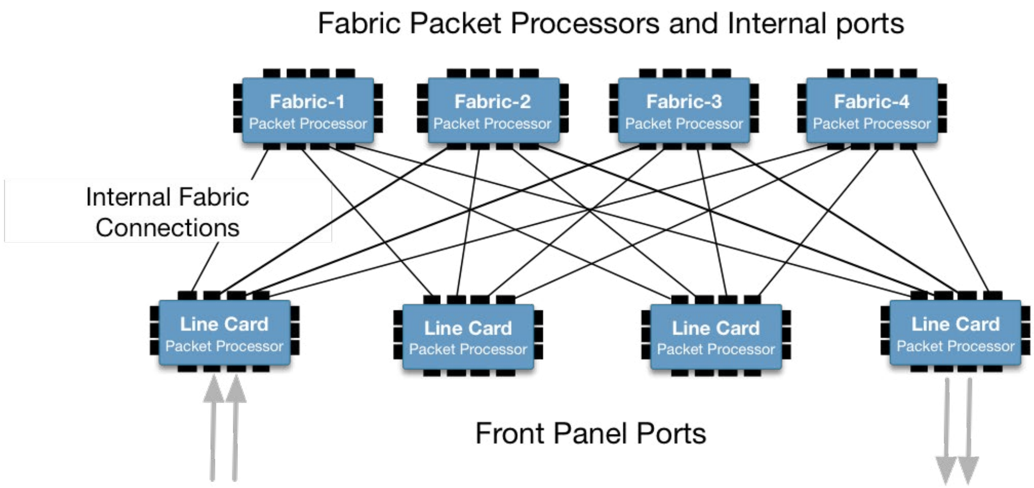

Figure 8: Internal Switching Fabric

Ethernet-Based Fabric

The first type of fabric is known as an Ethernet-Based fabric. As the name might suggest an Ethernet-Based fabric is largely bound by the rules of Ethernet. Consider a single chip as a switch connecting to other chips with patch cables all using Ethernet, an internal Clos design. Within an Ethernet-based design there are limits to the efficiency that can achieved. In general 80-90% efficiency is deemed achievable using bandwidth aware hashing and Dynamic Load Balancing (DLB) techniques on the linecard to fabric connections.

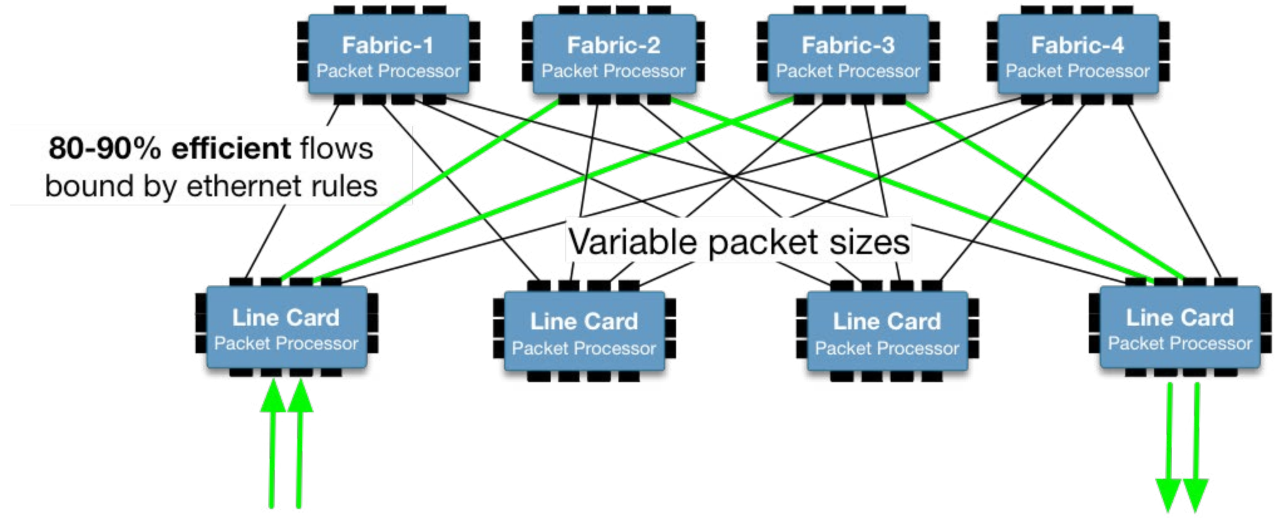

Figure 9: Ethernet-Based Fabric

Cell-Based Fabric - With Virtual Output Queuing (VOQ)

Cell based architectures are quite different as they are not bound by the rules of Ethernet on the switch backplane (the front panel port to fabric connections). A cell-based fabric takes a packet and breaks it apart into evenly sized cells before evenly "spraying" across all fabric modules. This spraying action has a number of positive attributes making for a very efficient internal switching fabric with an even balance of flows to each forwarding engine. Cell-based fabrics are considered to be 100% efficient irrespective of the traffic pattern.

Because the cell-based fabric does not utilize Ethernet it is inherently good at dealing with mixed speeds. A cell-based fabric is not concerned with the front panel connection speeds making mixing and matching 100M, 1G, 10G, 40G and 100G of little concern. Adding Advanced Queuing Credit based schedulers with Virtual Output Queues (VOQs) and deep buffers (for congestion handling) to a cell-based platform provides for a lossless based system that deserves consideration.

Cell based systems will give you more predictable performance under moderate to heavy load, the addition of Virtual Output Queue (VOQ) architectures will also help protect against packet loss during congestion. These two capabilities coupled with a deep buffer platform all but guarantee the lossless delivery of packets in a congested network.

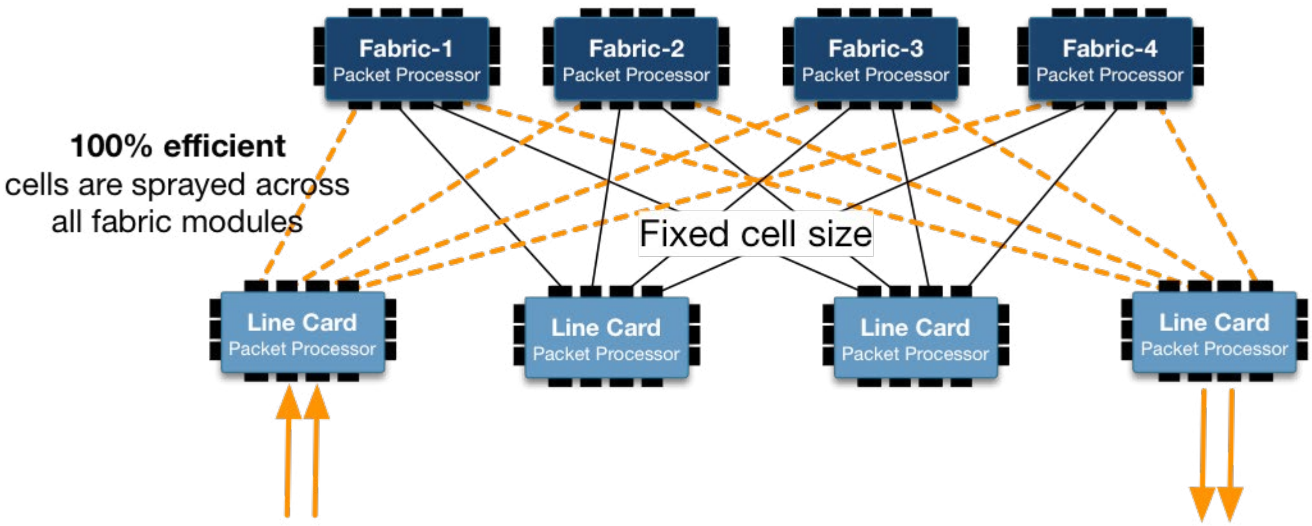

Figure 10: Cell-Based Fabric with VOQ

Choosing a Spine Platform

Like many design choices it comes down to having good data to work with. Some conditions to consider are: low loads, heavy loads, loads during failure conditions (loss of a spine) and loads during maintenance windows when a spine may be taken out of service. Ideally having good baselines is a good start, for net new builds this often comes down to modeling and predicated application demands. A proper capacity-planning program is essential to day two operations, ensuring your design can absorb or scale to meet future demands.

Here are some general rules of thumb that can help with the decision-making if you don't have enough data up front.

- In typical networks the first collision is seen at ~30% utilization
- Excessive collisions begin to happen at around ~70% utilization
- Networks begin to collapse above ~70% utilization if something isn't done

If you can't get a handle on the current and/or future design details, cell based large buffer systems are a catch all that makes the most sense when trying to minimize risk.

Hardware Specifications

The level of redundancy built in the spine as whole will dictate the level redundancy required at the platform level. A key consideration is that of supervisor redundancy. In a multi-spine design, the spines ability to lose a node without adversely impacting performance increases. By in large, multi-spine designs are configured to use a single supervisor since redundancy is provided by the other spine.

Leaf-Spine Interconnects

All leaf switches are directly connected to all spine switches through fiber optic or copper cabling. In an L2LS topology all of these interconnections are layer 2 links. These layer 2 links are bundled into Port-Channels and distributed across the two spine switches in a multi-chassis link aggregation group. Leaf-Spine interconnects require careful consideration to ensure uplinks are not over-subscribed. Subscription ratios can be engineered (as discussed in this document) to ensure uplinks are properly sized to prevent congestion and packet loss. Leaf-Spine Interconnects can be 10G, 40G or 100G interfaces.

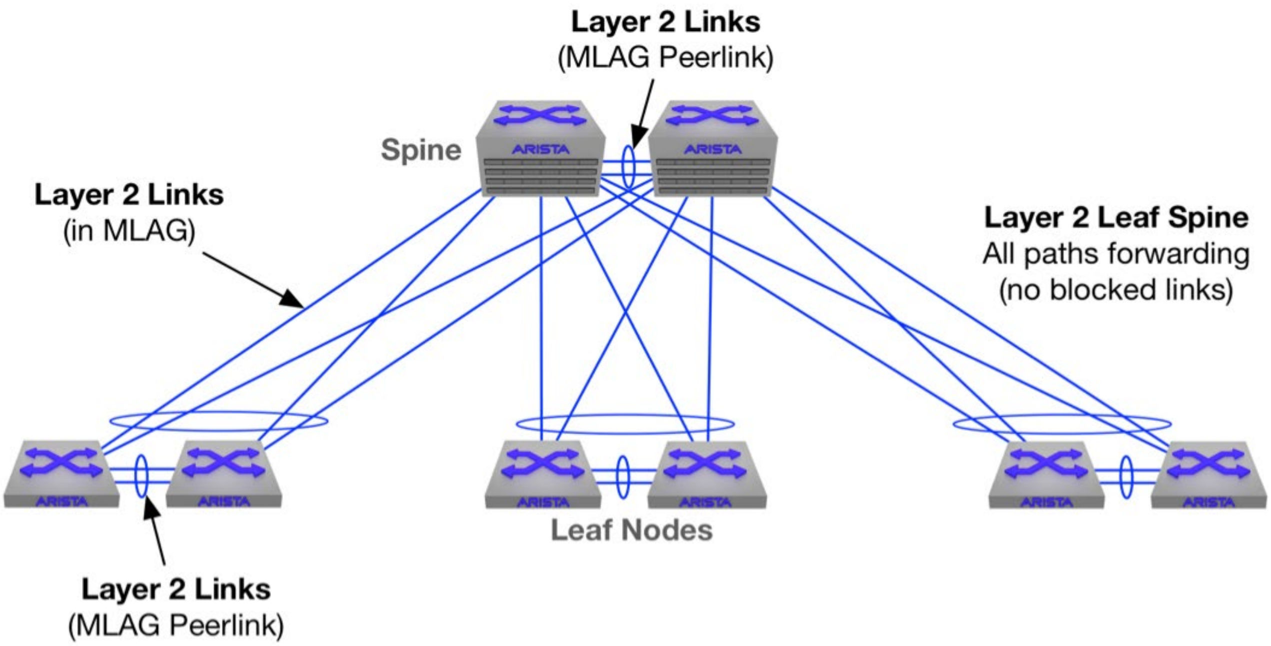

Figure 11: Layer 2 Leaf Spine Links

Congestion Management

Subscription Ratios

Subscription and over-subscription ratios are expressed as a ratio of downlink to uplink capacity. An example of this would be a 48 Port 10G switch with four 40G uplinks. In this scenario the over subscription ratio would be 480G:160G or 3:1. Oversubscription can exist in the North-South direction (traffic entering/leaving a data center) as well as East-West (traffic between devices in the data center).

For this design servers/workloads are attaching to the leaf at 1/10/25/50G. This is important because the bandwidth each server demands is aggregated when it comes to upstream connectivity i.e. the bandwidth consumed on the uplinks. Even though it's possible to have a wirespeed switch it does not mean servers will not encounter congestion. Server virtualization further compounds this problem as virtual machines (workloads) can pop up anywhere at anytime with no need for physical cabling. To ensure servers, both physical and virtual, do not suffer packet loss due to congestion, subscription ratios need to be taken into consideration. Below are some general rules of thumb when it comes to subscription ratios.

- Subscription-ratio of 1:1 for racks hosting IP Storage
- Subscription-ratio of 3:1 for General computing racks
- Subscription-ratio of 6:1 for racks containing Network services
- Subscription-ratio of 1:1 for Edge routing (match external BW with Spine BW)

When calculating subscription ratios it is important to know a little more about the quantity of servers (both virtual and physical) connecting to leaf switches as well as the expected bandwidth requirements. With 48 1/10G ports, there is a maximum of 480G of data coming into (ingress) and out (egress) of the switch. Using a subscription ratio of 3:1 can determine the uplink capacity required, in this example 480G / 3 = 160G. A switch with four 40G uplinks can meet this requirement (4x40G = 160G).

With the introduction of 100G uplinks the oversubscription level can be reduced in the 1/10G general computing deployment to 1.2:1. In this example, 48 1/10G host facing ports and at least 4 100G uplinks towards the spine, the oversubscription is 1.2:1 (480G / 1.2 = 400G).

Some common examples of host facing port to uplink ratios are outlined in the table below with options for 1/10G, 25G and 50G.

Table 1: Leaf subscription ratio examples

Host Ports	Uplinks	Over-Subscription	Example Hardware Configuration
48 1/10G	4 40G	3:1	48 1/10G (SFP+ or Copper) + 6 40G QSFP ports
48 1/10G	4 100G	1.2:1	48 1/10G (SFP+ or Copper) + 6 100G QSFP100 ports
48 1/10G	6 100G	1:1	48 1/10G (SFP+ or Copper) + 6 100G QSFP100 ports
96 25G [48 50G]	8 100G	3:1	96 25G [48 50G] (24 QSFP100) + 8 100G QSFP100 ports

Buffering

In general, congestion management and buffering does not seem to be well understood when it comes to data center networks. In reality, any network can experience congestion, and when it does buffers are utilized in an effort to avoid dropping packets. While a well thought out leaf and spine design minimizes oversubscription, services such as dedicated IP storage systems are prone to receiving large amounts of incast traffic. These types of traffic patterns have the potential to create bottlenecks.

Incast or TCP incast is a many to one communication pattern that is most commonly seen in environments that have adopted IP based storage as well as High Performance Computing applications as such as Hadoop. Incast can occur in different scenarios but a simple example is one where many hosts request data from a single server simultaneously. Imagine a server connected at 10G trying to serve 40G of data from 1000 users, this is a many to one relationship. Sustained traffic flows that exceed the capacity of a single link can cause network buffers to overflow causing the switch to drop packets.

For a detailed explanation on the benefits of buffers in the data center see the following white paper titled Why Big Data Needs Big Buffer Switches http://www.arista.com/assets/data/pdf/Whitepapers/BigDataBigBuffers-WP.pdf

When deep buffer systems are required Arista recommends the 7280 or 7500 series switches. In the 7500 and 7280 series systems, each port is capable of providing up to 50ms of packet buffering.

The implementation of a Leaf and Spine architecture with smaller buffers will perform well in a network that does not experience congestion. This makes it critically important to understand the performance characteristics required prior to making product decisions. Performing a good baseline analysis of traffic flows, particularly during periods of high utilization, will ensure your design meets your requirements.

Data Center Bridging and Priority Flow Control

Data Center Bridging (DCB) and Priority Flow Control (PFC) are two additional protocols used to assist with the lossless delivery of Ethernet.

DCB has two important features: Data Center Bridging Exchange (DCBX) and Priority Flow Control (PFC). EOS uses the Link Layer Discovery Protocol (LLDP) and the Data Center Bridging Capability Exchange (DCBX) protocol to help automate the configuration of Data Center Bridging (DCB) parameters, including the Priority-Based Flow Control (PFC) standard, which enables end-to-end flow-control.

As an example, these features enable a switch to recognize when it is connected to an iSCSI device and automatically configure the switch link parameters (such as PFC) to provide optimal support for that device. DCBX can be used to prioritize the handling of iSCSI traffic to help ensure that packets are not dropped or delayed.

PFC enables switches to implement flow-control measures for multiple classes of traffic. Switches and edge devices slow down traffic that causes congestion and allow other traffic on the same port to pass without restriction. Arista switches can drop less important traffic and tell other switches to pause specific traffic classes so that critical data is not dropped. This Quality of Service (QoS) capability eases congestion by ensuring that critical I/O (in the storage example) is not disrupted or corrupted and that other non-storage traffic that is tolerant of loss may be dropped.

Configuration Examples

This section of the guide is intended to provide configuration examples for all aspects of the layer 2 Leaf and Spine deployment. It includes sample configurations as well as steps to verify the configuration where appropriate. The following configuration examples will be covered.

- Base Configuration
- Management Interfaces
- MLAG Configuration (Spine / Leaf)
- Server Connectivity (VLANS and Gateways)

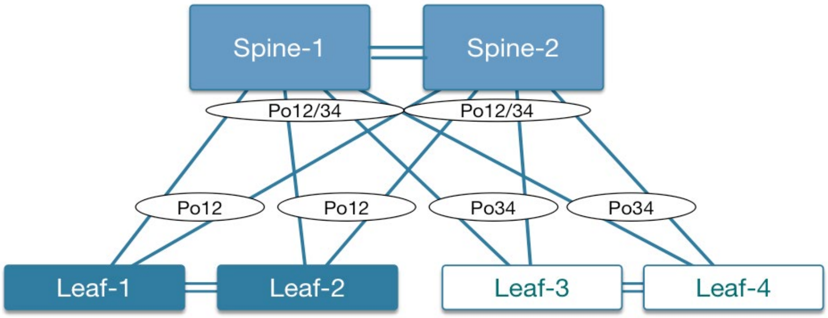

Figure 12: Topology for Configuration Examples

For a more thorough explanation of the specific commands found in this guide please see the Arista

System Configuration Guide for the version of EOS you are running. System Configuration Guides can be found on the Arista Support Site at http://www.arista.com/en/support/product-documentation.

Base Configuration (All Switches)

Below is a base configuration; hostname, DNS server addresses, Domain name, as well as NTP server information addresses are required before proceeding.

Spine-1 Example ! hostname Spine-1 ip name-server $DNSHostAddress ip name-server $DNSHostAddress ip domain-name $CompanyDomainName ! ntp source Management1/1 ntp server $NTPHostAddress1 prefer ntp server $NTPHostAddress2 ! username admin role network-admin secret 5 $Password ! clock timezone $tz	Notes: The variables below represent company/region specific information. $DNSHostAddress $DNSHostAddress $CompanyDomainName $NTPHostAddress1 $NTPHostAddress2 $Password $tz

Management Interfaces

Use Table 2 below as a reference to configure the management interfaces for all spine and leaf switches. The configurations for Spine-1 are shown below. Note that a VRF is used for management.

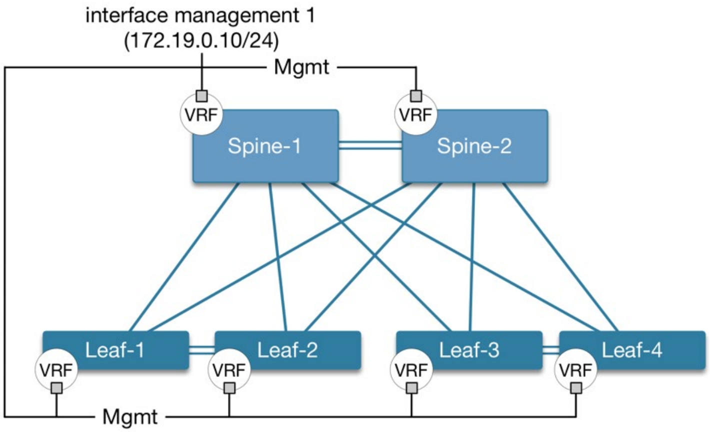

Figure 13: Management Interfaces

Spine-1 Example	Notes:
! vrf definition MGMT rd 0:65010 ! interface Management1 vrf forwarding MGMT ip address 172.19.0.10/24	This example is for Spine-01, see table x for the details for other spines.

Table 2: Management IP Addressing

Node	Interface	IP/Mask
Spine-1	Management 1/1	172.19.0.10/24
Spine-2	Management 1/1	172.19.0.11/24
Leaf-1	Management 1	172.19.0.12/24
Leaf-2	Management 1	172.19.0.13/24
Leaf-3	Management 1	172.19.0.14/24
Leaf-4	Management 1	172.19.0.15/24

MLAG Configuration (Bow-Tie Spine/Leaf)

This section of the configuration will assign the appropriate MLAG configuration to each spine and leaf switch. Each leaf switch will utilize four 40Gbps interfaces. Each of the four leaf uplinks will connect to a separate spine and each spine will have a connection to each leaf switch. The 40G interfaces are configured as LAYER 2(switched) point-to-point port-channel/trunks with logging enabled for changes in link status. Although there is no need for spanning tree in this design, as MLAG will utilize all available links, STP still runs in the background to prevent potential loops should a rogue switch be connected in a non-MLAG configuration.

We'll be utilizing what's referred to as a "bow-tie" MLAG design. This involves an MLAG forming from the spines to the leaf nodes, as well as the leaf nodes to the spines. The below example will be for just a single spine pair connecting to a single leaf pair. The same concept can be applied to additional leaf pairs.

MLAG Configuration Summary

1. Configure MLAG peer-link (port channel connecting the pair of leaf switches)
2. Create MLAG VLAN (LAYER 2and LAYER 3)
3. Define MLAG Domain (must match on both switches)
4. Configure MLAG member ports

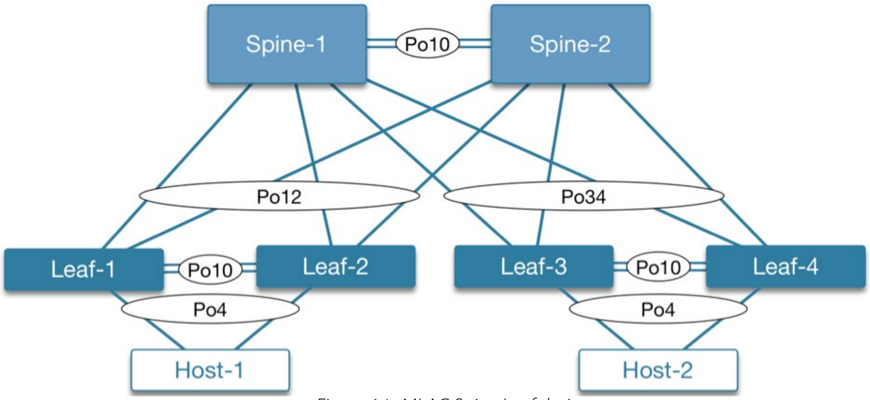

Figure 14: MLAG Spine Leaf design

Leaf Switches

Use the following MLAG sample configuration for Leaf-1 and Leaf-2.

Leaf-1 MLAG Configuration Example	Notes:
! vlan 4094 trunk group mlagpeer ! interface Port-Channel4 description Host-1 switchport access vlan 12 mlag 4 ! interface Port-Channel10 description MLAG Peerlink switchport mode trunk switchport trunk group mlagpeer ! interface Port-Channel12 description Leaf1-to-Spine switchport mode trunk mlag 12 ! interface Ethernet1 description MLAG Peerlink member port to Leaf2 mtu 9214 switchport mode trunk channel-group 10 mode active ! interface Ethernet2 description MLAG Peerlink member port to Leaf2 mtu 9214 switchport mode trunk channel-group 10 mode active ! interface Ethernet3 description TO SPINE1 mtu 9214 switchport mode trunk channel-group 12 mode active ! interface Ethernet4 description TO SPINE2 mtu 9214 switchport mode trunk channel-group 12 mode active ! interface Ethernet5 description Host-1 Link-1 mtu 9214 channel-group 4 mode active ! interface Vlan4094 description MLAG Peer Address Network ip address 172.16.12.1/30 ! mlag configuration domain-id mlag12 local-interface Vlan4094 peer-address 172.16.12.2 peer-link Port-Channel10	Configuration details are shown for Leaf-1 and Leaf2.

Leaf-2 MLAG Configuration Example

```
vlan 4094
  trunk group mlagpeer
!
interface Port-Channel4
  description Host-1
  switchport access vlan 12
  mlag 4
!
interface Port-Channel10
  description MLAG Peerlink
  switchport mode trunk
  switchport trunk group mlagpeer
!
interface Port-Channel12
  description Leaf2-to-Spine
  switchport mode trunk
  mlag 12
!
interface Port-Channel4
  description Host-1
  switchport access vlan 12
  mlag 4
!
interface Port-Channel10
  description MLAG Peerlink
  switchport mode trunk
  switchport trunk group mlagpeer
!
interface Port-Channel12
  description Leaf1-to-Spine
  switchport mode trunk
  mlag 12
!
interface Ethernet1
  description MLAG Peerlink member port to Leaf1
  mtu 9214
  switchport mode trunk
  channel-group 10 mode active
!
interface Ethernet2
  description MLAG Peerlink member port to Leaf1
  mtu 9214
  switchport mode trunk
  channel-group 10 mode active
!
interface Ethernet3
  description TO SPINE1
  mtu 9214
  switchport mode trunk
  channel-group 12 mode active
!
interface Ethernet4
  description TO SPINE2
  mtu 9214
  switchport mode trunk
  channel-group 12 mode active
!
interface Ethernet5
  description Host-1 Link-2
  mtu 9214
  channel-group 4 mode active
!
interface Vlan4094
  description MLAG PEER LINK
 ip address 172.16.12.2/30
!
mlag configuration
  domain-id mlag12
  local-interface Vlan4094
  peer-address 172.16.12.1
  peer-link Port-Channel10
```

NOTE: MLAG Configuration must be consistent on a pair of MLAG switches

Spine Switches

Use the following MLAG sample configuration for Spine-1 and Spine-2.

Spine-1 MLAG Configuration Example	Notes:

vlan 4094
  trunk group mlagpeer
!
interface Port-Channel10
  description MLAG Peerlink
  switchport mode trunk
  switchport trunk group mlagpeer
!
interface Port-Channel12
  description Spine-to-Leaf1-2
  switchport mode trunk
  mlag 12
!
interface Port-Channel34
  description Spine-to-Leaf3-4
  switchport mode trunk
  mlag 34
!
interface Ethernet1
  description MLAG Peerlink member port to Spine2
  mtu 9214
  switchport mode trunk
  channel-group 10 mode active
!
interface Ethernet2
  description MLAG Peerlink member port to Spine2
  mtu 9214
  switchport mode trunk
  channel-group 10 mode active
!
interface Ethernet3
  description TO LEAF1
  mtu 9214
  switchport mode trunk
  channel-group 12 mode active
!
interface Ethernet4
  description TO LEAF2
  mtu 9214
  switchport mode trunk
  channel-group 12 mode active
!
interface Ethernet5
  description TO LEAF3
  mtu 9214
  switchport mode trunk
  channel-group 34 mode active
!
interface Ethernet6
  description TO LEAF4
  mtu 9214
  switchport mode trunk
  channel-group 34 mode active
!
interface Vlan4094
  description MLAG PEER LINK
  ip address 172.16.11.1/30
!
	Configuration details are shown for Spine-1 and Spine-2.

```
  mlag configuration
  domain-id mlag01
  local-interface Vlan4094
  peer-address 172.16.11.2
  peer-link Port-Channel10
```

Spine-2 MLAG configuration example

```
vlan 4094
  trunk group mlagpeer
!
interface Port-Channel10
  description MLAG Peerlink
  switchport mode trunk
  switchport trunk group mlagpeer
!
interface Port-Channel12
  description Spine-to-Leaf1-2
  switchport mode trunk
  mlag 12
!
interface Port-Channel34
  description Spine-to-Leaf3-4
  switchport mode trunk
  mlag 34
!
interface Vlan4094
  description MLAG Peerlink
  ip address 172.16.11.2/30
!
mlag configuration
  domain-id mlag01
  local-interface Vlan4094
  peer-address 172.16.11.1
  peer-link Port-Channel10
!
interface Ethernet1
  description MLAG Peerlink member port to Spine1
  mtu 9214
  switchport mode trunk
  channel-group 10 mode active
!
interface Ethernet2
  description MLAG Peerlink member port to Spine1
  mtu 9214
  switchport mode trunk
  channel-group 10 mode active
!
interface Ethernet3
  description TO LEAF1
  mtu 9214
  switchport mode trunk
  channel-group 12 mode active
!
interface Ethernet4
  description TO LEAF2
  mtu 9214
  switchport mode trunk
  channel-group 12 mode active
!
interface Ethernet5
  description TO LEAF3
  mtu 9214
  switchport mode trunk
  channel-group 34 mode active
!
```

interface Ethernet6
  description TO LEAF4
  mtu 9214
  switchport mode trunk
  channel-group 34 mode active
!
interface Vlan4094
  description MLAG PEER LINK
  ip address 172.16.11.2/30
!
mlag configuration
  domain-id mlag01
  local-interface Vlan4094
  peer-address 172.16.11.1
  peer-link Port-Channel10	

Server Connectivity (VLANs and Gateways)

General configuration examples for connecting servers into the leaf switches are shown below. A more general-purpose single-homed compute configuration as well as a dual-homed (active/active) compute configuration is also shown. Note that server level configuration always needs to be reviewed as well, particularly with dual-homed active/active configurations.

Single-Homed Leaf Configuration

An example of VLAN and Gateway configuration for a single-homed compute leaf is shown below. Notice ports are configured as access ports and assigned a VLAN. A Switched Virtual Interface (SVI) is created for each VLAN, which acts as the default gateway for the host/workload. LLDP is also enabled on the switch to support learning of neighbor information.

Leaf-1 Single-Homed Server VLAN Configuration	Notes:

lldp run
!
vlan 100
  name SERVER-VLAN-172.20.0.0/24
!
vlan 200
  name SERVER-VLAN-172.30.0.0/24
!
interface Vlan100
  description SVI-FOR-VLAN-100
  mtu 9214
  ip address 172.20.0.1/24
  no shutdown
!
interface Vlan200
  description SVI-FOR-VLAN-200
  ip address 172.30.0.1/24
  no shutdown
!
interface Ethernet 21
  description Host-A
  switchport access vlan 100
  no snmp trap link-status
  storm-control broadcast level 1
  spanning-tree portfast
  spanning-tree bpduguard enable
  no shutdown
!
interface Ethernet 22
  description Host-B
  switchport access vlan 200
  no snmp trap link-status
  storm-control broadcast level 1
  spanning-tree portfast
  spanning-tree bpduguard enable
  no shutdown
	LLDP is enabled by default in Arista EOS

Dual-Homed Leaf Configuration

Below is the VLAN and port configuration for Leaf-1 and Leaf-2. Note that Leaf-1 and Leaf-2 are MLAG peers and will share a common configuration. Default gateway configuration uses the system VARP address. Both access and trunk port configurations are show. LLDP configuration is also shown in this example.

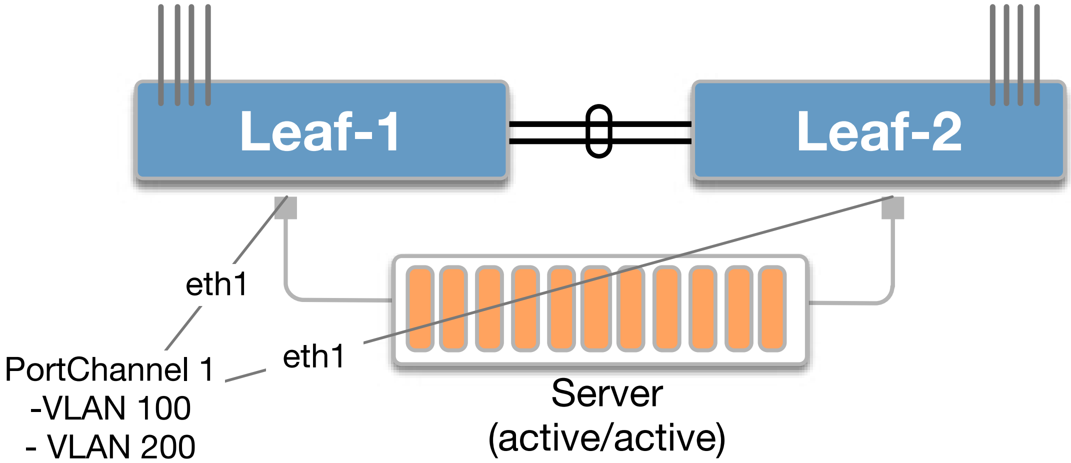

Figure 15: Dual-Homed Leaf Configuration for Server Connectivity

Leaf-1 Server VLAN Configuration

```
lldp run
!
vlan 100
  name SERVER-VLAN-172.20.0.0/24
!
vlan 200
  name SERVER-VLAN-172.30.0.0/24
!
interface Vlan100
  description SVI-FOR-VLAN-100
  mtu 9214
  ip address 172.20.0.2/24
  ip virtual-router address 172.20.0.1
  no shutdown
!
interface Vlan200
  description SVI-FOR-VLAN-200
  ip address 172.30.0.2/24
  ip virtual-router address 172.30.0.1
  no shutdown
!
interface Port-Channel1
  description PortChannel to Host-1
  switchport mode trunk
  switchport trunk vlan allowed 100,200
  no snmp trap link-status
  port-channel lacp fallback static
  mlag 1
  storm-control broadcast level 1
  spanning-tree portfast
  spanning-tree bpduguard enable
  no shutdown
!
interface Ethernet 1
  description MLAG-To-Server
  channel-group 1 mode active
  no shutdown
```

Leaf-2 Server VLAN Configuration

```
lldp run
!
vlan 100
  name SERVER-VLAN-172.20.0.0/24
!
vlan 200
  name SERVER-VLAN-172.30.0.0/24
!
interface Vlan100
  description SVI-FOR-VLAN-100
  mtu 9214
  ip address 172.20.0.3/24
  ip virtual-router address 172.20.0.1
  no shutdown
!
interface Vlan200
  description SVI-FOR-VLAN-200
  ip address 172.30.0.3/24
  ip virtual-router address 172.30.0.1
  no shutdown
!
interface Port-Channel1
  description PortChannel to Host-1
  switchport mode trunk
  switchport trunk vlan allowed 100,200
  no snmp trap link-status
  port-channel lacp fallback static
  mlag 1
  storm-control broadcast level 1
  spanning-tree portfast
  spanning-tree bpduguard enable
  no shutdown
!
interface Ethernet1
  description MLAG-To-Server
  channel-group 1 mode active
  no shutdown
```

List of Acronyms

IETF - Internet Engineering Task Force

LACP – Link Aggregation Control Protocol

IGMP – Internet Group Management Protocol

LLDP – Link Layer Discovery Protocol

L2 - Layer 2

L3 - Layer 3

MLAG - Multi-chassis Link Aggregation Group

PIM – Protocol Independent Multicast

RFC - Request for Comment

RIB - Routing Information Base

VARP - Virtual Address Resolution Protocol

VRRP – Virtual Router Redundancy Protocol

References

Arista Universal Cloud Network Design Guide http://arsta.co/2mGK3C1

Arista Cloud Networking Portfolio http://arsta.co/2pVsY9S

Arista Transceivers and Cables http://arsta.co/2qvTtQb

Arista EOS Configuration Guide http://arsta.co/2pVjz22

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office
1390 Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A & B, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062